On the role of statistics in the era of big data: a call for a debate

Piercesare Secchi

MOX-Department of Mathematics, Politecnico di Milano, Milano, Italy

Abstract

While discussing the plenary talk of Dunson (2016) at the 48th Scientific Meeting of the Italian Statistical Society, I formulated a few general questions on the role of statistics in the era of big data which stimulated an interesting debate. They are reported here with the aim of engaging a larger audience on an issue which promises to change radically our discipline and, more generally, science as we know it. But is it so?

Keywords: paradigm shift in statistics, big data analytics, high dimensional and complex data, two cultures, teaching statistics, distributed inference, data integration.

1. A sort of introduction

"Big data", is this the latest buzzword on the market or a credible indication of a paradigm shift which is changing science, not to mention statistics and the way we collect and analyze data? Without even a small answer to this big question, I was called to discuss the plenary talk *Probabilistic inference for big & complex data* delivered by David Dunson at the 48th Scientific Meeting of the Italian Statistical Society, held at the University of Salerno in June 2016. In the lack of profound insights, I thought I could at least formulate a few questions that would help me and my peers to size up some facets of the problem. These questions are here proposed again as an expedient to stimulate a discussion

on the role of statistics in the era of big data, if such an era exists and won't soon disappear as a feeble rhetorical invention. First, however, a disclaimer is in order: by no means my questions cover the entirety of the debate on big data, which is by now already very rich even on the pages of statistical journals. Indeed I hope that other and more knowledgeable discussants will fill the gap and touch upon aspects of the involved relationship between big data analytics and statistics which I left, consciously or not, underground.

One thing however seems indisputable, the big data trade is generating an humongous array of ad hoc analytics with a velocity, volume and variety competing with those that are said to capture the essence of the new data ecosystem. Be that as it may, unified theoretical frameworks for the statistical analysis - and inference - of big data are still missing, generating in the lay statistician the unspoken impression that we are dealing with alchemy, while chemistry is yet to come.

In *Pilgrim at Tinker Creek*, Annie Dillard (1974) wrote:

> If we are blinded by the darkness, we are also blinded by the light.
> When too much light falls on everything, a special terror results.

Too many unfiltered new ideas hamper innovation. This is the starting point of the latest book by Roberto Verganti (2017), *Overcrowded: Designing Meaningful Products in a World Awash With Ideas.* In the dark we can light a candle, but what shall we do when we are blinded by too many lights? Verganti's dictum is to pursue innovation driven by meaning, placing the human back at the center. This could as well be a precious indication for the statistician lost in the forest of "Big Data Analytics" under the spell of automatic science.

In the next section I will recall the questions which ignited the debate after Dunson's plenary talk at the 48th Scientific Meeting of SIS. Their original aim was to expose a few critical and general issues raised by the big data approach

to science, not just statistics. Indeed, although not all the questions are immediately related to the challenges posed to statistics by the current demand of analytics for huge datasets, these issues necessarily engage statistics as the practice dedicated, by tradition, to the collection and analysis of data. In the concluding section, I will however touch upon a few theoretical challenges which I believe have the potential to become important in a future dominated by the analysis of big data.

2. The seven questions at the SIS meeting

2.1. Is there a role for statistics in the big data era?

I would describe statistics as the science of variability, meaning that the main goal of statistics is to develop paradigms, methods and algorithms for the mathematical exploration, elicitation and control of variability, and the uncertainty it generates. Inference and uncertainty quantification are at the core of statistics and they have generated correlated siblings like prediction, testing, controlling for dependence, confounding, randomization. Yet these fundamental ideas of statistics are not often considered the primeval sources of the big data enlightenment. Is this the beginning of the end for statistics as we know it?

2.2. After the big data deluge, where do we stand in the debate about the two cultures in statistical modeling?

In 2001 Leo Breiman wrote on *Statistical Science*:

> There are two cultures in the use of statistical modeling to reach conclusions from data. One assumes that the data are generated by a given stochastic data model. The other uses algorithmic models and treats the data mechanism as unknown. The statistical community has been committed to the almost exclusive use of data models. This commitment has led to irrelevant theory, questionable conclusions, and has kept statisticians from working on a large range of

> interesting current problems. Algorithmic modeling, both in theory and practice, has developed rapidly in fields outside statistics. It can be used both on large complex data sets and as a more accurate and informative alternative to data modeling on smaller data sets. If our goal as a field is to use data to solve problems, then we need to move away from exclusive dependence on data models and adopt a more diverse set of tools.

After the big data deluge, where do we stand in the debate about the two cultures in statistical modeling?

2.3. Shall we abandon parametric statistics for good?

Big data are often characterized by huge sample sizes and a huge number of variables presenting very complex form of dependence. This is at the antipodes of the "received version" of statistics where datasets have small or moderate sample size and a reduced number of variables. In many big data applications, no parametric model is likely to capture the entire relevant variability occurring in the sample. Despite this, parametric models are still predominant both in the theory and in the practice of statistics. Is there a risk that in the big data era, parametric statistics will dangerously move from fitting models to data to fitting data to models? Could a truly non-parametric approach represent a point of contact between the two cultures evoked by Leo Breiman? Indeed in non-parametric statistics, on the one hand, one assumes that the data are generated by a stochastic model, and on the other one, the stochastic model is treated as (almost) completely unknown.

2.4. Not always massive, but often complex...

The big in big data is frequently a shorthand for massive in volume, velocity and variety. But data could also be big in complexity. More and more often, our statistical analyses involve data objects that are not easily reduced to a

Euclidean vector representation - the battlefield of classical Multivariate Data Analysis - without loosing the information content they support. I am thinking about applications of statistics in modern science where the atoms of the analysis are functions and surfaces, positive definite matrices or tensors, manifold data, trees, networks, texts. Object Oriented Data Analysis (Wang and Marron, 2007) provides a framework and a useful mode of discussion, for approaching complex data challenges. In the big data era what will be the data objects of the statistical analysis?

2.5. Teaching the next generation of statisticians

Good big data analyses require a deep understanding of the "physical" phenomenon generating the data. Whether we are Bayesians or not, this should be reflected in the prior knowledge which dictates the statistical paradigm and model, the optimization criteria (e.g shrinkage estimators or penalized regression), the sensitivity analysis, and, finally, the storytelling by which we communicate the results. Moreover data are often complex mathematical objects. Will this have an impact on teaching the next generation of students in statistics, which will move from being computer wise applied probabilists to knowledgeable data scientists (i.e. together with mathematics, a lot more of basic science - physics, chemistry, biology along with computer science and engineering, in the curricula of the next students in statistics)? What should be the role of advanced mathematics in the education of the next generation of statisticians?

2.6. Data visualization; the next frontier?

In his famous *Exploratory Data Analysis*, John Tukey (1977) wrote:

> The greatest value of a picture is when it forces us to notice what we never expected to see.

The high complexity of big data together with the multiplicity of sources generating them, challenge the traditional custom for the representation of their

information content and that provided by their statistical analysis, which up to now was conveyed through media like large tables of numbers or two-dimensional diagrams. The challenge is twofold. First, we need new visualization approaches for the representation of extremely complex data with minor losses of useful information, in an interactive, multi-dimensional, and multi-scale way which will enhance our potential for data exploration. Second, we need new tools for the visual communication of the results drawn from complex data analyses that rely on an extremely advanced knowledge in mathematics and statistics, ultimately enabling scientists, stakeholders, and policy makers to truly question big data. Are we statisticians open to the legacy of John Tukey (1977) and Edward Tufte (1983) and ready to take up the big data visualization challenge, working shoulder to shoulder with experts from those distant fields like communication design and visual literacy?

2.7. Reproducibility in the Big Data era?

David Donoho in 2010 wrote:

> An article about computational result is advertising, not scholarship. The actual scholarship is the full software environment, code and data, that produced the result.

The general problem of reproducibility has recently generated a strong debate, mostly conducted in top scientific journals:

> No research paper can ever be considered to be the final word, but there are too many that do not stand up to further study (*Nature*, 2017).

What guidelines shall we set for the papers published in our top journals in statistics and concerned with the analysis of big data? Can we statisticians claim to be totally clean when it comes to the peer-reviewed publishing system of our own journals?

3. Conclusion, or is it the begininning of the debate?

To be sure the statistical analysis of very big datasets, as well as that of high dimensional and complex data, challenges traditional approaches and raises new methodological issues. Dunson (2016) discusses some recent and ongoing schemes for scaling up probabilistic and Bayesian methods for big data settings, focusing on new efficient algorithms which modify existing MCMC for computing posterior distributions when the sample size is very large and the dimensionality of the data is comparatively small. Efron (2010, 2016) indicates that for "very large n, small p" problems, a renewed Empirical Bayes perspective promises to gain momentum, by allowing for the decoupling of the prior distribution of the unobserved parameter from an empirical estimate of the marginal distribution of the observables (see also the recent monograph by Efron and Hastie (2016)). On the complex data front, infinite dimensional curves and surfaces are the object of Functional Data Analysis, a term coined by Jim Ramsay in the 80's. FDA lately developed into a very rich and stimulating area of theoretical and applied research in data science, after the publication of the book by Ramsay and Silverman (1997) - by now already a classic - followed by that of Ferraty and Vieu (2006). Shapes (Dryden and Mardia, 1998), data on manifolds, trees and, more generally, data far from the familiar Euclidean setting are the focus of the already cited Object Oriented Data Analysis (Wang and Marron, 2007), recently reviewed by Marron and Alonso (2014).

Two connected topics which require fresh methodological efforts and are closer to my current line of research are distributed inference and data integration. They are pillar topics of the new Center for Analysis Decision and Society within the Human Technopole (`https://www.htechnopole.it/en/`), the research infrastructure established by the Italian Government and to be developed at the former Expo site in Milan. The Human Techonopole intensive

cross-disciplinary project is aimed to the synergistic development of fundamental and clinical genomics, innovative algorithms for data analysis, multiscale methods in computational life sciences, and advanced technologies for diagnostics, with the mission of making headway in personalized approaches, both medical and nutritional, focusing on cancer and neurodegenerative diseases.

Distributed data storage and processing is often considered "no more than" a data management system problem. However, the analysis of massive datasets has also stimulated a host of *divide et impera*, perturbative approaches to inference, where the original data sample is split in subsamples, each subsample is statistically analyzed in parallel and the results of these analyses are then aggregated in a final global one. Sometimes a *divide et impera* approach is a necessity due to computational or storage restrictions, but there are interesting situations where this line of action is recommended from a modeling perspective. For instance, when data - e.g. complex and high dimensional data - are spatially dependent and it is reasonable to assume only local stationarity for the random field generating the observations. An approach based on random domain decompositions, like that proposed in Menafoglio et al. (2017), whose predecessor is the Bagging Voronoi Algorithm (Secchi et al., 2013, 2015), can then be efficiently used for prediction, regression or classification. The inferential properties of these algorithms are still to be explored as well as a proper quantification of uncertainty, based on the decoupling of the endogenous variability - due to the stochastic process generating the data and to the sampling design - from the exogenous variability introduced by the random sub-sampling. Indeed, even in the iid case, a naive implementation of distributed statistical inference might lead to suboptimal estimators; see, for instance, Jordan et al. (2016), Tang et al. (2016), Banerjee et al. (2016) and the literature cited therein.

Similarly, data integration is not only an information technology problem.

Loosely stated, the statistical problem of combining information from two distinct databases is that of formulating useful models for inferring the joint distribution of the random elements (X, Y), from the data contained in one database and generated by the distribution of the random elements (W, X), and the data contained in a distinct database and generated by the distribution of the random elements (W, Y). The easy way out is to assume conditional independence of X and Y, given W, and then estimate the joint distribution of (W, X, Y), but this is often an unrealistic mathematical escape. Moreover the random elements X or Y could be complex and high dimensional (e.g. object data); realistic models for their conditional distribution are not available. The problem is not new, and indeed, within a Gaussian multivariate framework, it is already considered in Kadane (2001), a reprint of a manuscript dated 1978. See also the more recent book by D'Orazio et al. (2006). As a referee of this paper pointed out, problems of data integration and data fusion come into several different flavors. Given the easy availability of massive information relative to the same population, one problem is that of combining at a microlevel different and heterogeneous data sources, which only partially share variables, like when combining census data with open data collected from different administrative databases and survey data. A somewhat different problem of data integration appears in geospatial applications, when data are sampled along different spatial grids and time frames of the same space-time domain. Or the data sources are relative to spatial decompositions belonging to different levels - like when combining demographic and socioeconomic data collected at the county level, state level, national level ... - and, moreover, the analysis requires to integrate these data with those generated by phenomena, like climate or natural risks, whose spatial dependence is captured by models which hardly conform to the spatial domain decompositions implied by the former data sources. As it happens

life sciences create a huge amount of data and information stored in different databases; if properly exploited, managed, and integrated, this information and data may drive the evolution of epidemiology, clinical studies, and personalized medicine. However, new, and computationally efficient, methods are needed for integrating the heterogeneous datasets which are increasingly necessary to answer large-scale questions concerning, for example, citizen wellbeing and disease progression. Analogous data integration methods could also find applications in the setting up of integrated platforms for natural risks prevention and reduction. This is for instance the case of the plan *Casa Italia* promoted by the Italian Government after the tragic earthquake of Amatrice in August 2016. With an integrated, multi-hazard perspective, Casa Italia aims at identifying the several natural and societal risks that jointly plague the Italian territory (e.g seismic, volcanic, hydrogeological,...) and to devise policies which act for the mitigation and containment of vulnerability and exposition, when it is not possible to intervene directly on hazard reduction. It is therefore necessary to integrate very heterogeneous data generated by different agencies, having different space and time references and recorded with different degrees of uncertainty : census and demographic data, seismic hazard estimates based on historical data, microzonation data, flood hazard assessments based on mathematical models of basins, fragility curves and indicators for buildings based on their dimensions, age and construction material, economic indicators...

Some (Anderson, 2008) claim that the advent of big data is marking the end of science as we know it - a model based testable approach to the understanding of reality - while enhancing what is asserted to be a truly orthodox Baconian perspective, by automating the scientific process through an intensive use of computers and a blinded search for correlations. I disagree and I suspect mystification. Small or big, data don't speak for themselves and correlation is

not enough. And yet I believe that the challenges posed by big data will fuel not only a renewed market interest in the data analysis ecosystem, but also new theory and methods in statistics, whether the new name of statistics will be data science or not. If only statisticians will vigorously rise to the challenge and shake off the misgivings about big data being only a catchy term coined by computer scientists.

4. Acknowledgements

Most of the thoughts, issues and perspectives reported in the previous pages are the result of fruitful discussions with my fellows of the Stat Group at MOX (`https://statistics.mox.polimi.it`). I thank them all with the hope that they will forgive me if I have misrepresented their views on the subject of statistics in the big data era.

References

Anderson C. (2008). The end of theory: the data deluge makes the scientific method obsolete. *Wired*, `https://www.wired.com/2008/06/pb-theory/` (last access: 05.17.2017)

Banerjee M., Durot C. and Sen B. (2016). Divide and Conquer in Non-standard Problems and the Super-efficiency Phenomenon. *arXiv:1605.04446v3*

Breiman L. (2001). Statistical modeling: the two cultures. *Statistical Science*, 6(3), 199 231.

D'Orazio M., Di Zio M. and Scanu M. (2006). *Statistical Matching: theory and practice*, Chichester: Wiley.

Dillard A. (1974). *Pilgrim at Tinker Creek*. New York : Harper's Magazine Press.

Donoho D. L. (2010), An invitation to reproducible computational research, *Biostatistics*, 11(3), 385-388

Dryden, I. L. and Mardia, K. V. (1998). *Statistical Analysis of Shape*. Wiley, New York, NY.

Dunson D. (2016). Probabilistic inference for big & complex data. *48th Scientific Meeting of the Italian Statistical Society*, June 8-10 2106, Università degli Studi di Salerno, Italy.

Efron B. (2010). *Large scale inference: Empirical Bayes methods for Estimation, Testing and Prediction*. Institute of Mathematical Statistics Monographs, vol.1, Cambridge University press.

Efron B. (2016). Empirical Bayes Deconvolution methods. *Biometrika*, 101(1), 1–20.

Efron B. and Hastie T. (2016). *Computer age statistical inference*. Institute of Mathematical Statistics Monographs, vol.5, Cambridge University press.

Ferraty F. and Vieu P. (2006). *Nonparametric functional data analysis: theory and practice*. Springer Verlag, New York.

Jordan M. I., Lee J.D. and Yang Y. (2016). Communication-Efficient Distributed Statistical Inference. *arXiv:1605.07689v3*

Kadane J.B. (2001). Some statistical problems in in merging data files. *Journal of Official Statistics*, 17, 423-433.

Marron J. S. and Alonso A. M. (2014). Overview of object oriented data analysis. *Biometrical Journal*, 56 (5), 732753.

Menafoglio A., Gaetani G. and Secchi P. (2017). Random Domain Decompositions for object-oriented Kriging over complex domains. *Manuscript.*

Nature (2017). Editorial of the Special on Challenges in Irreproducible Research. `http://www.nature.com/news/reproducibility-1.17552\#/Editorial` (last access: 05.17.2017)

Ramsay J. O. and Silverman B. W. (1997), *Functional Data Analysis.* Springer-Verlag, Berlin; New York.

Secchi P., Vantini S. and Vitelli V. (2013). Bagging Voronoi classifiers for clustering spatial functional data. *International Journal of Applied Earth Observation and Geoinformation*, vol. 22, p. 53-64.

Secchi P., Vantini S. and Vitelli V. (2015). Analysis of spatio-temporal mobile phone data: a case study in the metropolitan area of Milan (with discussion), *Statistical Methods and Applications*, 24(2), 279-300

Tang L., Zhou L. and *Song P.* (2016). Method of Divide-and-Combine in Regularised Generalised Linear Models for Big Data. *arXiv:1611.06208*

Tukey J. W. (1977), *Exploratory Data Analysis.* Reading, Mass: Addison-Wesley Pub.

Tufte E. (1983), *The visual display of quantitative information.* Cheshire, CT: Graphics Press.

Verganti R. (2017). *Overcrowded: Designing Meaningful Products in a World Awash With Ideas.* Cambridge: The MIT Press.

Wang H. and Marron J.S. (2007), Object oriented data analysis: sets of trees, *Annals of Statistics*, 35(5), 1849-1873